AN EMPIRICAL ANALYSIS OF THE VALUE OF COMPLETE INFORMATION FOR eCRM MODELS

By: Balaji Padmanabhan
Operations and Information Management Department
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6366
U.S.A.
balaji@wharton.upenn.edu

Zhiqiang Zheng
A. Gary Anderson School of Management
University of California, Riverside
Riverside, CA 92521
U.S.A.
eric.zheng@ucr.edu

Steven O. Kimbrough
Operations and Information Management Department
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6366
U.S.A.
kimbrough@wharton.upenn.edu

Abstract

Due to the vast amount of user data tracked online, the use of data-based analytical methods is becoming increasingly common for e-businesses. Recently the term analytical eCRM has been used to refer to the use of such methods in the online world. A characteristic of most of the current approaches in eCRM is that they use data collected about users’ activities at a single site only and, as we argue in this paper, this can present an incomplete picture of user activity. However, it is possible to obtain a complete picture of user activity from across-site data on users. Such data is expensive, but can be obtained by firms directly from their users or from market data vendors. A critical question is whether such data is worth obtaining, an issue that little prior research has addressed. In this paper, using a data mining approach, we present an empirical analysis of the modeling benefits that can be obtained by having complete information. Our results suggest that the magnitudes of gains that can be obtained from complete data range from a few percentage points to 50 percent, depending on the problem for which it is used and the performance metrics considered. Qualitatively we find that variables related to customer loyalty and browsing intensity are particularly important and these variables are difficult to derive from data collected at a single site. More importantly, we find that a firm has to collect a reasonably large amount of complete data before any benefits can be reaped and caution against acquiring too little data.

Keywords: Data mining, incomplete data, information value, eCRM

Veda Storey was the accepting senior editor for this paper. Chrysanthos Dellarocas, Maytal Saar-Tsechansky, and Young Rhu served as reviewers. The associate editor chose to remain anonymous.
Introduction

Effective customer relationship management (CRM) is important for any business (Pan and Lee 2003, Varian 2001). It has been shown that customer data can play a critical role in designing effective CRM strategies (Padmanabhan and Tuzhilin 2003). This is particularly the case in the online world where Web servers automatically collect vast amounts of behavioral data on customers. The term analytical eCRM (Swift 2000) has been used to refer to the use of data-based analytical methods for customer analysis, customer interactions, and profitability management. A characteristic of most of the existing approaches to eCRM is that they build profiles and models based on data collected by a single Web site from users’ interactions with the site. In this paper we refer to such data as site-centric data, which we define to be clickstream data collected at a site augmented with user demographics and cookies (Sen et al. 1998). In this sense, traditional approaches are myopic: they are based on firms building models from data collected at their site only. However, the myopic nature of most current eCRM methods is not due to the fact that site-centric data is adequate for understanding customer behavior; rather, it is due to the nature of data ownership: most sites only have access to their own log files.

For example, consider two users who browse the Web for air tickets. Assume that the first user’s session is as follows: Cheaptickets1, Cheaptickets2, Travelocity1, Travelocity2, Expedia1, Expedia2, Travelocity3, Travelocity4, Expedia3, Expedia4, Cheaptickets5, where Xi represents some page i, at Website X and in this session assume that the user purchases a ticket at Cheaptickets. Assume that the second user’s session is Expedia1, Expedia2, Expedia3, Expedia4, and that this user purchases a ticket at Expedia (in the booking page Expedia, in particular). Expedia’s (site-centric) data would include the following:

User1: Expedia1, Expedia2, Expedia3
User2: Expedia1, Expedia2, Expedia3, Expedia4

We define user-centric data to be site-centric data plus data on where else the user went in the current session. In this sense, user-centric data is the “complete” version of site-centric data. In the above example, the user centric data for Expedia is:

User1: Cheaptickets1, Cheaptickets2, Travelocity1, Travelocity2, Expedia1, Expedia2, Travelocity3, Expedia3, Travelocity4, Expedia4, Cheaptickets5
User2: Expedia1, Expedia2, Expedia3, Expedia4

Reconsider the site-centric data for Expedia. In one case (user 2), the first three pages result in the user making a purchase at the next page. In the other case (user 1), the first three pages result in no purchase. Expedia sees the “same” initial browsing behavior, but with opposite results. However sophisticated the analytical methods used, it is difficult to differentiate these two sessions based on site-centric data alone. Nevertheless, most conventional techniques implicitly try to do so, coerced by incomplete information.

The example above suggests that user-centric data may be more informative than site-centric data for problems in eCRM. This observation is indeed actionable since technologies exist to collect user-centric data. Firms can obtain user-centric data in one of two ways.

1. Firms can get customers to download client-side tracking software that captures a user’s online activities. Clearly in this case, firms also have to provide appropriate incentives for users and should be explicit to their users about what data is tracked.

2. Market data vendors, such as Netratings and comScore Networks, collect user-centric data by signing up panels and using client-side tracking software to capture all of the browsing activities of their panelists. Firms can purchase user-centric data directly from such vendors.

A firm’s decision to acquire such data should be based on whether the benefits exceed the costs of obtaining the data. In this paper, we focus on quantifying the gains that may be obtained from user-centric data, and on quantifying how much user-centric data is necessary to obtain to outperform site-centric data. By doing so, this paper provides answers that will help firms in understanding the benefits that may be obtained from having complete information. Actual decisions, of course, will also depend on the particulars of the individual businesses.

Specifically, we consider three important and common problems in eCRM and for each problem we evaluate how user-centric models compare to site-centric models. For each of these three problems, we answer the following questions:

1. What is the magnitude of the gain that can be expected from acquiring complete user-centric data, and what factors drive the gains that are obtained? This is an important question since the answers to this can help understand the benefits of user-centric data. However, from a practical perspective, although there may be gains from user-centric data, cost considerations alone may prevent firms from being able to acquire user-centric data.
from all their customers. This motivates the second question that we study.

2. If it is only possible to acquire user-centric data for a sample of customers, how much data should be acquired in order to do better than with site-centric data alone? We use the term critical mass to refer to this number. In addition to the practical reason for studying this question, this is interesting from a research perspective since it addresses a tradeoff between having more (less informative) site-centric data or less (more informative) user-centric data.

The main results are that the magnitudes of the gains are significant, but the actual values depend on the performance metric. User-centric data can increase the predictive accuracy of the desired target variable by 10 to 20 percent for the three problems considered. The gains are more striking for computing lift gains, where the magnitude of the gains is between 20 and 50 percent. Further, analyses of the models built reveal several important predictors. In particular, the key predictors from user-centric data capture effects that relate to customer loyalty and browsing intensity—both of which are hard, if not impossible, to determine accurately from site-centric data. Further, the critical mass results reveal that the critical mass numbers are fairly high, ranging from 25 to 45 percent depending on the problem and metric considered. This is an important result since it suggests that if firms acquire partial user-centric data, then acquiring too little can be counter-productive in that user-centric models built from such data can actually be worse than site-centric models.

The rest of this paper is organized as follows. In the next section, we explain our choice of the eCRM problems considered, the modeling technique used, and the performance metrics of the models (for which the gains are studied). Based on this discussion we state the goals of the paper more specifically in terms of determining two 3 × 3 tables. We then present the methodology used and, in particular, elaborate data preparation algorithms. The results are presented, followed by a discussion and conclusions.

Specific Objectives

In order to compare models built on site-centric data to models built on user-centric data, three dimensions have to be clearly specified: (1) the choice of the problems for which the models are built, (2) the choice of the modeling technique used, and (3) how the comparison of site-centric and user-centric models is done. In this section, we develop reasonable choices for these three dimensions, and based on this discussion we identify the specific values that need to be estimated.

Choice of the Problems

In order to compare models built on site-centric and user-centric data, we consider the following important and common problems in eCRM:

• Predicting repeat visits of users. This is important in the context of understanding loyalty and switching propensities of online users. In particular, we build models that, at the end of each user session, predict if the user is likely to visit a site again at any time in the future. In prior work, Lee (2000) modeled repeat visit behavior of online users using an NBD (negative binomial distribution) model. Chen and Hitt (2002) studied consumers’ switching behavior between online brokers based on repeat visits using a logit model.

• Predicting likelihood of purchase at a site. Within this category we identify two subproblems. The first is a real-time prediction task: within a current user’s session, at every successive click, the task is to predict if this user is likely to purchase within the remainder of the session. We term this the within-session prediction problem. The second is a prediction problem with a longer time horizon: at the end of each user session, predicting if this user is likely to purchase during any subsequent session in the future, the future-session prediction problem. In prior work, VanderMeer et al. (2000) developed a personalization system to predict a user’s next access in the session, and issue signals when the user is going to make a purchase. Theusinger and Huber (2000) modeled online users’ within-session purchase of software. Moe and Fader (2004b) studied users’ future-session purchase behavior for Amazon.com. Fader and Hardie (2001) investigated customers’ future-session purchase behavior for CDNow and discovered that customers’ purchase propensity diminishes over time.

Choice of the Models

Prior work for these problems built a wide range of different models ranging from linear models (Mena 1999), logit models (Chen and Hitt 2002), other probabilistic models (Lee 2000; Moe and Fader 2004a, 2004b) and nonlinear models (e.g., classification trees and neural networks—see Theusinger and
Huber 2000; VanderMeer et al. 2000). In our experiments we build classification tree models for the following reasons:

1. All three tasks considered are binary and classification trees have been shown to be highly accurate for such tasks (Padmanabhan et al. 2001).
2. This approach makes few distributional assumptions about the data and can accommodate a wide range of data types.
3. In our experiments we build site-centric and user-centric models for 95 sites and three problems. For a single model choice, this amounts to $2 \times 95 \times 3 = 570$ models. A single neural network that we built took approximately 2 hours on a Linux workstation with 1G main memory and a 1GHz processor. Classification tree algorithms are fast and, therefore, enable comparison on a larger scale.
4. Finally, classification trees enable visual interpretation of the results and provide useful variable importance reports.

**Model Comparison Methods**

In our experiments, when models on site-centric and user-centric data are compared, we focus on out-of-sample performance of these models. In particular, for each of the 570 datasets considered, the data is split so that a random 40 percent is used as in-sample to build the model and the remaining 60 percent is used as out-of-sample data on which the model is evaluated. In the out-of-sample data we compute three metrics for each model: (1) the overall predictive accuracy, (2) the target variable predictive accuracy, and (3) a measure of how much “lift” the model provides.

A default metric for comparing approaches is overall predictive accuracy. However, when the dependent variable has unequal priors, this method has limitations (Fawcett and Provost 1996, Johnson and Wichern 1998). For example, assume that 5 percent of the sessions result in a purchase. A trivial rule that there will be no purchases in any session will have an accuracy of 95 percent. However, this is clearly not useful. Part of the problem is that, in such cases, failure to recognize a potential purchase could result in losing the customer, and it is therefore more expensive to label a booking session as a non-booking one. One method of addressing this limitation is to specify prior misclassification costs and build models to minimize total costs (Johnson and Wichern 1998). The problem with this method is that it is sensitive to the actual misclassification costs used and in many applications it may not be easy to determine these costs and, therefore, the choice of these costs may be ad hoc. In light of this, we consider two additional metrics. First, besides overall prediction accuracy, we consider the target variable’s predictive accuracy (i.e., what percentage of the actual purchases were predicted correctly). For example, the trivial model stated above that had 95 percent overall accuracy will have 0 percent target variable accuracy. Second, as is standard in the database marketing literature (Hughes 1996; Ling and Li 1998), we use a lift index to compare models.

Assume that a model predicts that the $i$th record in the out-of-sample data is a “booking” session with confidence $p_i$. The data is then sorted in descending order of $p_i$ and a lift curve is constructed as follows. Any point $(x, y)$ belongs on the lift curve if the top $x$ percent of this sorted data captures $y$ percent of the actual booking sessions. If the model is random the top $x$ percent of the data would be expected to contain $x$ percent of the bookings. The difference $(y - x)$ is hence the lift obtained as a result of the model. For example, the model in Figure 1 picks out 50 percent of the true booking sessions from just the top 20 percent of the sorted data. The area between the 45-degree line and the lift curve of the model is a measure of how much “lift” this model provides and we use the value of this area as the index of the lift. Clearly this is within the range $[-0.5, 0.5]$ and the greater the value, the better is the model.

Hence in this paper our goal is to compute two $3 \times 3$ tables in which the rows are the specific problems and the columns are the performance metrics. The first table represents the magnitude of the gains that can be expected from having complete user-centric data. The other table represents the critical mass numbers (i.e., the percentage of user-centric data that is needed to build models that are as good as the models built from the entire site-centric data). In the next section, we describe the methodology used to determine these values.

**Methodology**

The gains and critical mass numbers may be obtained as follows.

1. Select $N$ sites and obtain site-centric and user-centric data for each site.
2. Prepare each site’s user-centric and site-centric datasets (construct relevant variables and apply feature selection) as applicable for each problem and thereby obtain $N$ pairs of datasets for each problem.
3. For each problem, build N pairs of models and for each performance metric, determine the overall magnitude gains and critical mass averaged over the N observations.

Since step 3 is straightforward, in this section we describe how the first two steps are done.

**Generating Site-Centric and User-Centric Data**

Ideally we would like to select N Web sites and obtain site-centric and user-centric data directly from them. This is impossible since no site on the Web has complete user-centric data on its customers and site-centric data is usually proprietary. However, we could reproduce various sites’ site-centric and user-centric data if we had a random panel of users and observed their behavior across several sites. Firms such as Jupiter Media Metrix and Netratings collected such data at a user-level based on a large, randomly selected panel of users. In our research, we use raw data from Jupiter Media Metrix that consisted of records of 20,000 users’ Web surfing behaviors over a period of 6 months. The data included user demographics and transaction history over the entire period. The total size of the raw data amounted to 30GB and represented approximately 4 million total user sessions. The data that are captured in each user session are the set of visited pages and the times of visit. In addition, the vendor categorized the various sites accessed by the users into categories such as books, search engines, news, CDs, travel, etc. In particular, we chose 95 sites spanning five categories (book, music, travel, auction, and general shopping malls) that represent sites that sell products such that each site had at least 500 user sessions. Below we present more formally the algorithms to create site-centric and user-centric data from such raw data.

First, we present the algorithm used for constructing each site’s site-centric data from raw user-level browsing data. Figure 2 presents CalcSiteData, an algorithm for constructing site-centric data from user-level Web browsing data.

Let $S_1, S_2, ..., S_N$ be N user sessions and $usr$ be a function such that $usr(S_i)$ represents the user in session $S_i$. We define each session $S_i$ to be a tuple $(p_{i1}, p_{i2}, ..., p_{ik})$, where $p_{ij}$ is the $j$th page accessed in session $S_i$ and $ki$ is the length of session $S_i$. We say a page $p$ is in session $S_i$ if $p$ is any element in $(p_{i1}, p_{i2}, ..., p_{ik})$. We also define an operation `add_page` such that `add_page(p_{1}, p_{2}, ..., p_{k}, q)` returns $(p_{1}, p_{2}, ..., p_{k}, q)$. Finally, let function `rootdomain(p)` return the Website’s root domain (e.g., `www.expedia.com`) of a given page $p$ (e.g., `www.expedia.com/xyz.html`).

CalcSiteData works by taking each user session and constructing snapshots for each unique site in the session such that the snapshot consists of pages belonging to that particular site. For example, given a session `<Cheapickets, Cheaptickets, Travelocity, Expedia, Expedia, Travelocity, Expedia, Expedia, Cheaptickets,>`, CalcSiteData extracts three tuples:

1. `<Cheapickets, Cheaptickets, Cheaptickets,>` for site `Cheapickets`
2. `<Travelocity, Travelocity, Travelocity, Travelocity,>` for site `Travelocity`
3. `<Expedia, Expedia, Expedia,>` for site `Expedia`
Inputs: (a) User sessions $S_1, S_2, ..., S_n$, (b) functions $usr$ and $rootdomain$

Outputs: Site-centric data for each Website site.

for (i = 1 to N) {
    Let $UniqSites = \{X |$ page p in $S_i$ such that $X = rootdomain(p)\}$
    for all (site $\in UniqSites$) {
        $site_centric_session = <>$
        for all pages p, in session $S_i$ {
            if (rootdomain(p) = site) {
                $site_centric_session =$
                add_page($site_centric_session, p)$
            }
        }
        output 'site, demographics of $usr(S_i)$, site_centric session'
    }
}

Figure 2. Algorithm CalcSiteData

Inputs: (a) User sessions $S_1, S_2, ..., S_n$, (b) functions $usr$ and $rootdomain$

Outputs: User-centric data for each Website site.

for (i = 1 to N) {
    Let $UniqSites = \{X |$ page p in $S_i$ such that $X = rootdomain(p)\}$
    for all (site $\in UniqSites$) {
        output 'site, demographics of $usr(S_i)$, $S_i$'
    }
}

Figure 3. Algorithm CalcUserData

From the tuples extracted from all of the user sessions, grouping the tuples for each individual site results in the site-centric data for that site. In addition to the tuples representing the individual pages, site-centric data also contains user demographics contained in the raw data. CalcSiteData is presented in Figure 2.

Observe that what CalcSiteData tries to do is to recreate each site’s Web logfile data based on the sample of 20,000 users’ Web accesses. Given that it is impossible to obtain the complete logfile data for every commercial site, the strength of CalcSiteData is that it can simulate (and adduce) individual logfiles from publicly available user-level browsing data.

The creation of user-centric data for each site is straightforward. For each site $s$, the subset of all user sessions that contain $s$ represents the user-centric data of the site. For instance, the user session <Cheaptickets, Cheaptickets, Travelocity, Travelocity, Expedia, Expedia, Travelocity, Travelocity, Expedia, Cheaptickets> will belong to the user-centric data of Expedia, Travelocity, and Cheaptickets. Algorithm CalcUserData is formally presented in Figure 3.

Data Preparation for the Three Problems

We first summarize the data preparation method for the within-session prediction problem. The other two problems (future-session prediction and repeat visit prediction) share a common structure: at the end of a session predicting a binary event in the future. The data preparation for these two problems is hence the same and we discuss this second.
**Data Preparation for the Within-Session Problem**

The goal of the model being built is to predict at any point within a session if a purchase will occur in the remainder of the session. Consider a user’s session \(<p_1, p_2, p_3, p_4, p_5>\) at a site where a purchase occurs in page \(p_5\). This single session is not a single data record for modeling. Rather, it provides five data records. Three sessions that began with \(\{p_1, p_2, p_3\}\) and \(\{p_1, p_2, p_3\}\) respectively resulted in a subsequent purchase. Two that began with \(\{p_1, p_2, p_3, p_4, p_5\}\) did not.

This distinction is important. It indicates that sessions create data records proportional to their length. In general, a session of length \(k\) provides \(k\) data records for modeling. In total, the number of records is, therefore, the sum of all session lengths. Given the explosion in the number of points this creates we sample each session probabilistically based on its length. For example, if a sampling rate is 0.2, a session of length \(k\) on average provides \(0.2 \times k\) data records for modeling. Associated with each sampling, a random “clipping point” is chosen within the session. Explanatory variables constructed from information before the clipping point are used to predict whether a purchase occurs in the part of the session after the clipping point.

A key strength of probabilistic sampling is that the size of the preprocessed data can be chosen based on time and space constraints available. Choosing a maximum desired data size, \(dnm\), is equivalent to choosing a sampling rate per session of \(dnm/numtotal\), where \(numtotal\) is the sum of the lengths of all the sessions in the data. The actual algorithm, probabilistic clipping, is presented at Appendix A.

Preprocessing user-centric data for the within-session prediction problem is similar to the method described above. This involves using probabilistic clipping partially, that is, only for creating fragments of all sessions in site-centric data and then augmenting these fragments with user-centric information before creating the summary variables. For example, consider a single user’s session. Let the site-centric and user-centric data for Expedia contain the following records:

- **Site-centric record**: \(\text{Expedia}_1, \text{Expedia}_2, \text{Expedia}_3, \ldots, \text{Expedia}_n\)
- **User-centric record**: \(\text{Cheaptickets}_1, \text{Cheaptickets}_2, \ldots, \text{Cheaptickets}_n, \text{Travelocity}_1, \text{Travelocity}_2, \ldots, \text{Travelocity}_n, \text{Expedia}_1, \text{Expedia}_2, \ldots, \text{Expedia}_n, \text{Travelocity}_1, \text{Travelocity}_2, \ldots, \text{Travelocity}_n, \text{Expedia}_1, \text{Expedia}_2, \ldots, \text{Expedia}_n\)

If the sampling rate is 0.7, assuming that as part of the probabilistic clipping procedure for site-centric data, the following two fragments of the session are created \(\text{Expedia}_1, \ldots, \text{Expedia}_3\) and \(\text{Expedia}_4, \ldots, \text{Expedia}_7\). Preprocessing user-centric data for Expedia would then create the fragments \(\text{Cheaptickets}_1, \ldots, \text{Cheaptickets}_3, \text{Travelocity}_1, \ldots, \text{Travelocity}_2, \text{Expedia}_1, \ldots, \text{Expedia}_3, \text{Cheaptickets}_4, \ldots, \text{Cheaptickets}_7, \text{Travelocity}_4, \ldots, \text{Travelocity}_7, \text{Expedia}_4, \ldots, \text{Expedia}_7\). Based on these two fragments, summary variables are created. Clearly the summary variables in user-centric data will contain additional metrics such as percentage of total hits to Expedia’s site in the current session (20 percent for the first fragment and 33 percent for the second fragment in the example). The detailed set of user-centric metrics and site-centric metrics is presented in Appendix B.

Note that as a result of probabilistic clipping, a session might result in several data records. This occurs since the prediction problem addressed here is to predict at every click whether a purchase is likely to occur. The data generated from probabilistic sampling is, therefore, the correct sample since we are sampling from the space of all session fragments. Since longer sessions have more session fragments, they will translate into more records. This is a natural result of what we want to sample from, and is not over-representation.

**Data Preparation for the Future-Session and Repeat Visit Problems**

The future-session (or repeat visit) prediction problem is predicting at the end of a given session, whether a user at a site will make a purchase (or visit the site again) at some point in the future. To illustrate the difference in using site-centric data and user-centric data, consider a user at a given point in time with three Web sessions involving visiting Expedia. The prediction task involving site-centric data is to predict if a given user with \(n\) prior sessions with Expedia will book at some future session based on the historical sessions \(s_1, s_2, \ldots, s_n\), where each \(s_i\) is of the form \(\text{Expedia}_i, \text{Expedia}_j, \text{Expedia}_k\) for example. The prediction task involving user-centric data is to predict if a given user with \(n\) prior sessions with Expedia, will book at some future session based on the historical sessions \(u_1, u_2, \ldots, u_m\), where each \(u_i\) is of the form \(\text{Cheaptickets}_i, \text{Cheaptickets}_j, \text{Travelocity}_i, \text{Travelocity}_j, \text{Expedia}_i, \text{Expedia}_j, \text{Travelocity}_i, \text{Travelocity}_j, \text{Expedia}_i, \text{Expedia}_j\) for example. Hence for a user with \(N\) total sessions at Expedia, the site-centric and user-centric data would each contain \(N\) records, each with increasing historical contents. The individual summary variables are explained in the beginning of Appendix B.

---

5We use Jupiter Media Metrix’s heuristic that considers properties of secure-mode transactions to infer bookings. In their experiments using 10 commercial sites that provided them exact URLs of purchases, this heuristic identified 87 percent of the total purchases and had a false positive rate under 25 percent.
Table 1. Average Number of Variables Chosen by Feature Selection

<table>
<thead>
<tr>
<th></th>
<th>Site-Centric</th>
<th>User-Centric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within Session</td>
<td>7.6 (15)</td>
<td>15.9 (40)</td>
</tr>
<tr>
<td>Future Session</td>
<td>7.3 (12)</td>
<td>15.5 (30)</td>
</tr>
<tr>
<td>Repeat-Visit</td>
<td>6.2 (12)</td>
<td>12.9 (30)</td>
</tr>
</tbody>
</table>

Results

We started with user-level browsing data consisting of 20,000 users’ browsing behavior over a period of 6 months (provided to us by Jupiter Media Metrix). We constructed site-centric and user-centric datasets for 95 sites using the process described earlier. The number of records for each site ranged from 500 to 12,500 raw user sessions. Based on the generated site-centric and user-centric data, we then applied feature selection to filter out possibly irrelevant variables for each site. The feature selection problem is the problem of selecting a “best” subset of features (variables) from a finite space of features (Olafsson and Yang 2002). There are several feature selection techniques. For our experiments, we choose a wrapper algorithm implemented in Weka, a popular Java-based data mining package (Witten and Frank 1999).

Table 1 reports the average number of variables selected by the wrapper algorithm. Note that we apply feature selection to each dataset separately. This means that two site-centric datasets may actually differ in the set of final variables selected from the entire set of available variables. For example, for the within-session problem, the selected features for the site-centric data and for the user-centric data for a specific site were (gender, age, hsize, minutelh, hpsesslh, mpselllh, freqlh) and (age,income, hpsessgh, basket, single, bookgh, hitgc, minutshc), respectively (see Appendix B for detailed explanations of these variables). This ensures that the best model is built for each possible site-centric or user-centric dataset. The values in parenthesis in Table 1 are the number of variables of each dataset before feature selection. Note that the feature selection algorithm greatly reduced the number of variables (e.g., for the site-centric data of the within-session problem, the average number of variables selected across 95 sites is 7.6 versus 15 original variables).

Using only those variables selected by the feature selection algorithm, we then built 570 models (2 × 3 × 95; two model types (user-centric and site-centric), three predictive problems, and 95 sites). A summary of the results is reported below.

Computing the Magnitude Gains

For the within-session problem, we apply probabilistic clipping with a sampling rate of 0.2 to obtain site-centric and user-centric datasets for each of the 95 sites’ raw session datasets. The resulting 95 pairs of datasets contain between 1,000 and 40,000 records each. As discussed earlier, the datasets for the future-session problem and the repeat-visit problem have the same structure. The datasets for these two problems contained between 500 and 12,500 records.

Each of these 570 datasets (95 × 3 × 2) was divided into a random 40 percent for model building and a remaining 60 percent as out-of-sample data. Note that in the out-of-sample data for user-centric models, we assume we know the complete user-centric predictor variables. Although decision trees are accurate classifiers, they belong to a family of techniques that are unstable in the sense that a perturbation in the dataset can result in a different tree generated. Hence for each dataset, we generate five decision trees based on different random 40-60 data partitioning. All the results thus are based on the five-run average. The actual performance of all the models is summarized in Table 2, which reports the mean and variance for site-centric models and user-centric models for all three problems under three performance measures. In addition, we report the t-values and p-values from the paired t-tests between site-centric models and user-centric models. Based on the out-of-sample model comparisons for 95 sites, the user-centric models significantly outperform site-centric models in all cases (all nine p-values are less than 0.05).

Table 3 specifically lists the magnitude gains (in percentage) derived from Table 2. The results show that user-centric data can increase the predictive accuracy of the desired target variable by 10 percent to 20 percent for the three problems considered, while maintaining the overall accuracy. The gains are more pronounced for lift gains, where the magnitude of the gains is between 21 percent and 48 percent. As mentioned in the introduction, in this paper we empirically study the modeling gains that are obtained. Translating these gains into dollar terms is important yet is dependent on specific
Table 2. Summary Results

<table>
<thead>
<tr>
<th></th>
<th>Overall Accuracy</th>
<th>Target Variable Accuracy</th>
<th>Lift Gains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Variance</td>
<td>Mean</td>
</tr>
<tr>
<td>Within-Session</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Site-centric</td>
<td>0.854</td>
<td>0.005</td>
<td>0.425</td>
</tr>
<tr>
<td>User-centric</td>
<td>0.869</td>
<td>0.004</td>
<td>0.511</td>
</tr>
<tr>
<td>t</td>
<td>5.460</td>
<td></td>
<td>5.210</td>
</tr>
<tr>
<td>p</td>
<td>&lt; 0.001</td>
<td></td>
<td>0.003</td>
</tr>
<tr>
<td>Future-Session</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Site-centric</td>
<td>0.875</td>
<td>0.006</td>
<td>0.299</td>
</tr>
<tr>
<td>User-centric</td>
<td>0.883</td>
<td>0.007</td>
<td>0.337</td>
</tr>
<tr>
<td>t</td>
<td>3.517</td>
<td></td>
<td>3.086</td>
</tr>
<tr>
<td>p</td>
<td>&lt; 0.001</td>
<td></td>
<td>0.003</td>
</tr>
<tr>
<td>Repeat-Visits</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Site-centric</td>
<td>0.656</td>
<td>0.004</td>
<td>0.480</td>
</tr>
<tr>
<td>User-centric</td>
<td>0.671</td>
<td>0.004</td>
<td>0.537</td>
</tr>
<tr>
<td>t</td>
<td>3.503</td>
<td></td>
<td>5.753</td>
</tr>
<tr>
<td>p</td>
<td>&lt; 0.001</td>
<td></td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 3. The Magnitude of Gains (in Percentage)

<table>
<thead>
<tr>
<th></th>
<th>Overall Accuracy</th>
<th>Target Variable Accuracy</th>
<th>Lift Gains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-Session</td>
<td>1.8%</td>
<td>20.2%</td>
<td>48.6%</td>
</tr>
<tr>
<td>Future-Session</td>
<td>0.9%</td>
<td>12.7%</td>
<td>21.2%</td>
</tr>
<tr>
<td>Repeat-Visits</td>
<td>2.3%</td>
<td>11.9%</td>
<td>39.9%</td>
</tr>
</tbody>
</table>

decision contexts. It is beyond the scope of this paper to investigate the effect of these gains on specific decisions. However, to provide a more intuitive sense of the values, we describe how these numbers may be interpreted for the future session prediction problem.

Assume that at some specific point in time a firm desires to score all its online visitors based on how likely they are to purchase at a future session (perhaps to decide whom to target for some promotion). Further, assume that there are 5 million visitors to the site and that 1 million of them actually do purchase at a later time. Finally, assume that the expected performance measures of the user-centric and site-centric models are as reported in Table 2. Based on an overall accuracy gain of 0.01, the user-centric model would correctly predict the propensity to purchase for 50,000 (i.e., 5 million × (0.88 - 0.87)) more users than the site-centric model. Perhaps more important, among the users who actually make a purchase, the user-centric model would correctly identify 40,000 (i.e., 1 million × (0.33 - 0.29)) more purchasers than the site-centric model. Finally, note that the user-centric model has an increase in lift of 0.024 compared to the site-centric model. For some loyalty promotion, assume that the firm wishes to target the 500,000 most likely purchasers. Both user-centric models and site-centric models would first be applied to sort the users from most likely purchasers to least likely purchasers, and then the best 500,000 users would be selected for the promotion. Although it is impossible to exactly determine the gains without actually doing this sorting, roughly assuming that the lift of the user-centric model is uniformly higher by the same percentage throughout the lift curve suggests that the user-centric model will identify 24,000 more purchasers (2.4% × 1 million) than the site-centric model would from this sorted list.

Identifying the Drivers of the Gains

Figure 4 indicates that in our data, 85 percent of user sessions involve the user visiting more than one site, with 27 percent...
also visiting more than 10 Web sites. Given these numbers, if the additional user-centric data is relevant, then adding this information can help build better predictive models.

Also, the models from user-centric data clearly indicate that in addition to the effects captured by site-centric models, there are several other factors, not captured in site-centric data that have highly significant effects. Due to space considerations, we present only one pair of sample trees with the top three levels.

Figure 5 represents the sample trees of Dell built from site-centric data and user-centric data for the future-session problem. The site-centric tree is relatively simple with only two levels that represent three rules. In contrast, all if the variables in the top three levels in the user-centric tree are user-centric variables. Average number of sessions per site (sespsite), bookings in the past at any site (bookgh), the market share of a user’s time spent at this site (minutesh), and the percentage of sessions that start with this site (entrate) are all very significant for the future-session purchase prediction problem.

In order to determine what specific variables in user-centric data drive the results, we sorted the variables based on the following heuristic. For each tree (of the user-centric models) we consider the variables that appear in the splits of the top five levels. We weigh each level linearly, with variables in the first level given a weight 5, the second level a weight 4, and so on. We then extract the variables in the top 5 levels and add their importance for all 95 sites for each of the 3 problems. The 10 most important variables are reported in Table 4. Overall, 20 out of the 30 most important variables are user-centric ones. Detailed analyses below further bring out several important effects captured only in the user-centric case.

Figure 4. Distribution of Number of Sites Visited Across User-Sessions

Figure 5. Site-Centric Tree (Left) Versus User-Centric Tree (Right) of Dell Dataset
For the within-session prediction problem, 7 out of the 10 most important variables in the sorted list were user-centric variables that cannot be generated from site-centric data. The three most important ones, with strong effects, were (1) the percentage of historical sessions in which the current site was the one where the user spent the most time (peakrate) and (2) whether in the current session this site is the entry site or the one in which the user spent the most time (path) and (3) the percentage of sessions at this site in which the user stopped immediately after this site (exitrate). All of these three variables are based on the traversal path of a user within and across sites. Also note that these variables all capture some measures of user loyalty to a specific site. Inherently loyalty is extremely difficult to capture from site-centric data alone; this can never give you information on a user’s activity at competitors’ sites.

For the future session prediction problem, 6 out of the 10 most significant variables in the sorted list were user-centric measures. The three most significant user-centric variables in the list were (1) bookgh (past bookings at all sites), (2) peakrate (explained above), and (3) exitrate (explained above). The first is a measure of browsing/buying intensity, and not surprisingly heavy buyers are more likely to book in the future. The other two again capture loyalty effects. These findings indicate that loyalty is a significant driver of purchase behavior and one that can be better captured in user-centric data.

For the repeat visit problem, the most significant variable is a site-centric one, sesslh, the number of sessions of the user to this site in the past. Not surprisingly, this is a highly positive, though trivial, indicator of repeat visits. The next two are user-centric variables that capture browsing/buying intensity (number of sessions per site, sespsite, and total purchases in the past across all sites, bookgh). Among the most significant user-centric variables were basket size (the number of sites a user was visiting in the current session and awareset) and total number of unique shopping sites the user visited. Both variables are indicators of comparison shopping, and hence also capture loyalty related measures.

In summary, variables that capture two important drivers—loyalty and browsing/buying intensity—tend to be highly important for all three problems considered in this paper. Among these two, loyalty cannot be measured from the more common site-centric data. The other effect of browsing/buying intensity, although partially captured in site-centric data, is important enough that the actual values captured in user-centric data turn out to be much better.

### Critical Mass Results

Thus far we computed the magnitude of gains that can be obtained from having user-centric data for all customers. However, cost considerations may preclude a firm from

---

3 There is an interesting connection that may be made here to literature on decision making. According to Kahneman (1994), individuals heavily weigh the peak moment during an episode, and also strongly attend to how the episode ended. Also, Tversky and Kahneman (1974) found that decisions are characterized by significant reliance on the starting point.
acquiring all of the user-centric data. In order to determine how much user-centric data is needed, here we compute the critical mass numbers for all three problems and all three performance metrics. We define critical mass as the percentage of user-centric data that is required to build a model that is as good as the site-centric model. This metric is computed as $K_{\text{min}}/N$ where $K_{\text{min}}$ is the minimum amount of $N$ user-centric records needed to build a model that can “outperform” (in terms of the performance metrics) the site-centric model built from the $N$ site-centric records.

One way to compute the metric is by comparing the two scenarios shown in Figure 6. Scenario 1 represents model building with site-centric data, where we have all $N$ customers’ records yet with fewer variables ($X_1, \ldots, X_M$). In contrast, scenario 2 represents a user-centric model where we have fewer customers’ records ($K$ as shown in Figure 6) yet more variables (the additional user-centric variables $X_{M+1}, \ldots, X_p$ in Figure 6). The critical mass $K$ thus represents the tradeoff between knowing more customers (scenario 1) and knowing more information per customer (scenario 2).

Determining the critical mass based on comparing the two scenarios above is implemented as follows. For each site, we first build a site-centric model using all $N$ in-sample records available. We then randomly select $K$ records from the corresponding $N$ user-centric records and build a tree on the data. In particular, we vary the actual value $K$ from 2 to 100 percent incrementally, with step size at 2 percent. The critical mass is determined when the user-centric model with $K$ records starts to outperform the site-centric model.

Figure 7 graphically depicts the determination of the critical mass for an average site ubid.com$^4$ for the within-session problem under the “target variable accuracy” performance metric. In Figure 7, the performance of site-centric data is represented by the straight line (0.401 in terms of booking accuracy) since we use all $N$ available records. As $K$ increases, note that the performance of the user-centric model (shown as the dotted curve in Figure 7) increases as would be expected. The point at which the user-centric model outperforms the site-centric model afterwards is 44 percent. Therefore, the critical mass for ubid is 44 percent.

Table 5 summarizes the critical mass results$^5$ (mean and variance) for the three problems under the three chosen performance measures. All the critical masses are in the range of 25

---

$^4$Its critical mass of 44 percent is about the same as the average critical mass (43.9 percent) of the 95 sites.

$^5$All of the results are averaged across five runs with different random seeds when choosing $K$ random user-centric data records.
to 45 percent. This implies that for these problems and performance metrics, it would be necessary to acquire a reasonably large percentage of user-centric data in order to do better than existing site-centric data alone. This is an important result, since it suggests that too small a sample may actually be counter-productive in the sense that the user-centric models can indeed be worse than the site-centric models. Note as well the specific shape of the user-centric curve. Knowing this will facilitate estimating the amount of user-centric data required for a given level of improvement in the measure of performance, here booking accuracy, as part of a cost-benefit analysis.

We would like to note that the critical mass can be lowered by perhaps using better models. One way of lowering the critical mass is by combining site-centric data with the acquired user-centric data to build models. For example, noting that scenario 2 effectively ignores available site-centric data (on all customers for whom additional data was not acquired), one possibility is to use an imputation model built on the acquired user-centric data (scenario 2) to impute the missing values for all of the remaining site-centric data, and thereby build a user-centric model based on all customers. We implemented this approach by using SAS’s multiple imputation method and Table 6 reports the summary results. As the comparison shows, combining data appears to be a promising method to lower the critical mass of data needed, since for two out of the three metrics it does lower the critical mass a little for all three problems. For lift gains, the combined approach is comparable with the previous approach based on the critical mass of data needed. Perhaps with a more advanced combining method (e.g., using intelligent weights as a function of user-centric data available), the critical mass results could improve further. This is an interesting issue to study in future work.

### Discussion

The results presented in this paper are significant for the following reasons:

1. In the context of predictive modeling, even though the nature of the result is not surprising, there has been no
Table 6. Summary Critical Mass Results Using Combined Data

<table>
<thead>
<tr>
<th></th>
<th>Overall Accuracy</th>
<th>Target Variable Accuracy</th>
<th>Lift Gains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Variance</td>
<td>Mean</td>
</tr>
<tr>
<td>Within-Session</td>
<td>30.0%</td>
<td>0.235</td>
<td>38.6%</td>
</tr>
<tr>
<td>Future-Session</td>
<td>38.0%</td>
<td>0.235</td>
<td>40.4%</td>
</tr>
<tr>
<td>Repeat-Visit</td>
<td>36.7%</td>
<td>0.227</td>
<td>43.9%</td>
</tr>
</tbody>
</table>

2. Prior work that made this comparison explicitly to understand the magnitude of the gains that can be achieved from user-centric data. As expected, although the gains are significant, the magnitude of the difference between site-centric and user-centric approaches varies by problem type and performance metric. An interesting observation from Table 3 is that, for the within-session prediction problem, the gains that can be achieved are higher than for the other problems. It is worth noting that this is the most difficult one among the three problems and is one that most online retailers care about since being able to predict purchase (or the lack of it) within a session is often actionable. In terms of the performance metric, the gains appear larger for the lift metric but these gains for the different metrics cannot be easily compared.

2. For the specific problems considered, we determined the significant variables that impact the models. In particular, we find that variables capturing customer loyalty and browsing/buying intensity are significant. These variables can be computed better from user-centric data and to some extent help understand why user-centric data can be useful.

3. As discussed in the introduction, it may be appealing for firms to collect user-centric data on only some customers. For the specific problems and models considered in this paper, we show that too little user-centric data can actually be worse than just using the site-centric models. Our results here actually make a case for the use of site-centric data when it may not be possible to acquire the reasonably large amount of user-centric data needed. While this result is certainly not contradictory to the claim that complete user-centric data is better, it provides some degree of support for the common use of site-centric models in eCRM problems since for many firms the acquisition of a large amount of user-centric data may be prohibitively expensive.

4. There are interesting implications. These results suggest that there may be value in business models that collect user-level data and provide these to individual sites to build models. Indeed some potential opportunities include customer opt-in models for licensing user-level data real-time to electronic commerce sites for building more effective models. More generally, the methodology used in this paper may be useful in valuing additional user data and may also be useful in formally studying the value of privacy. These are exciting possibilities for future research.

There are several limitations that might limit the scope of the paper. First, the models do not consider the content within the pages since most were dynamically generated (e.g., a profile of an air ticket search) and cannot be recovered. However, note that user-centric data is a superset of site-centric data; there is no reason to believe that this will help site-centric models more than user-centric ones. Second, reproducing site-centric datasets from user-level panels is only an approximation. Given the impracticality of obtaining site-centric datasets directly from hundreds of individual sites, it is not possible to determine how good the approximation really is. Third, the results are specific to the variables, the metrics, and the model (decision tree) considered in this paper. In Appendix C, we discuss how the features were generated and present the mapping between user-centric and site-centric variables used in this study to point out where we could not generate site-centric variables. One of the interesting directions for future work is to ask what additional information can be augmented to site-centric models to make their performance closer to user-centric models, perhaps by acquiring additional user data or by using current scripting technologies to gather more Web browsing information automatically.

Future Research and Conclusions

There are several exciting opportunities for future research that extend this work. Below we describe three important extensions.
First, as mentioned in the introduction, future research is needed to determine the cost-benefit tradeoffs from a firm’s perspective. While there are different methodologies that can potentially be used to study this question, below we outline an analytical approach that can determine how much data to acquire. Let \( K \) be the number of customers to acquire. Assume a linear cost function\(^6\) of data acquisition \( C(K) = \alpha K \). As Figure 7 shows, the model accuracy is roughly a concave function of \( K \). Suppose there is a concave profit function \( P(K) \). Then the manager’s decision is an optimization problem to determine \( K^* \) such that the overall profit \( P(K) - \alpha K \) is maximized, subject to \( N > K > CM \), where \( CM \) is the critical mass. From the first order condition, we know that \( K^* \) satisfies \( P'(K^*) = \alpha \). If \( K^* \leq CM \), then the site is better off just using site-centric data to build the model. If \( CM < K^* < N \), then acquire \( K^* \) number of customers. If \( K \geq N \), then acquire all customers. Figure 8 illustrates this simple decision making scenario. This paper helps such an approach in that it can be used to determine (1) the critical mass and (2) the accuracies of the site-centric and user-centric models as a function of \( K \) (and these functions may then be used to develop an appropriate profit function).

Second, the observation that user-centric data may be better suggests that information integration strategies are worth studying, but future research needs to carefully weigh the benefits of information sharing with the costs incurred by all parties in doing so (including a potential reduction in benefits if all parties have full information). Information integration, in the online world, can be pursued at three very different levels, and we describe these possibilities below. Associated with each possibility are important research questions that need to be addressed in future work in order to build effective information integration strategies.

1. Firm–Firm: At this level, information integration is achieved by strategic alliances between firms to share data. Similar to ours, recent work in marketing (Park and Fader 2005), finance (Kallberg and Udell 2003), and economics (Japelli and Pagano 2002) demonstrates the value of information sharing at a firm level. Park and Fader investigated the benefits of information sharing between Amazon and Barnesandnoble. Kallberg and Udell studied the effect of lenders sharing credit information about borrowers and show that information sharing adds value in solving the problem of borrower information opacity. Japelli and Pagano find that bank lending is higher and default rates are lower, in countries where lenders share information, regardless of the private or public nature of the information sharing mechanism. Using a data mining based approach, our research complements the recent work in information sharing and is the first to demonstrate the value of this in the online world. However, an obstacle for interfirrm information sharing has been the need to maintain the privacy of user data. While current work in various disciplines argues for information sharing between firms, much research needs to be done to study market mechanisms and incentives to make this happen while being sensitive to privacy concerns in the online world.

2. Firm–Customer: At this level, information integration can be achieved by firms directly acquiring additional information from customers. Indeed this is much easier to do in the online world and can be enabled by using automated Web services, technologies that are getting...
much attention for their seamless applications integration capabilities. In a recent work, Zheng and Padmanabhan (2002) proposed an approach based on firms selectively choosing customers from whom to acquire additional data. The advantage of this approach is that privacy issues are automatically handled since customers choose to provide this data themselves. Hagel and Rayport (1997) predict that customers will put aside concerns over privacy if they receive sufficient value in return for providing personal information. One approach is to pay consumers for the use of their personal information, either through royalty payment, or by providing discounts as some supermarket loyalty programs do for their members. An open research issue is how to structure the right incentives for customers to reveal their private data in the online world.

3. Firm–Market: At this level, information integration can be achieved by obtaining information from a market that provides integrated customer information. This has interesting implications for user privacy since an information market can be an efficient way to deal with customers’ private data. Laudon (1996) proposes an innovative market-based solution to user privacy, in which individuals would have a common law property right to their personal information. These rights could be sold and national information markets would emerge. Such information markets can be easily created for the online world, and could be another source from which firms acquire customer information.

Third, as a by-product of our methodology, a number of significant predictor variables have been identified from the user-centric data that are not available from the site-centric data. (For examples, see Table 4 and Appendix B.) This presents the prospect of using data on these variables, along with other data, broadly for purposes of marketing, much as demographic data (ZIP code, occupation, etc.) are used.

In summary, in this paper we compared models using site-centric data versus models using user-centric data for three different prediction tasks common in eCRM applications. Based on these comparisons, we empirically determined the magnitude of the gains that can be achieved using user-centric data and also determined the minimum amount of user-centric data that needs to be acquired if a firm cannot acquire all user-centric data. We discussed the implications of these results and also outlined interesting questions for future research. The main contributions of the paper are the empirical determination of the gains, the specific user-centric variables identified as relevant, and the critical mass needed for each of the key problems considered in the paper. We also wish to note that the methodology used in the paper to answer the questions posed is novel and is an important contribution in itself. This methodology is one that can be applied to any user-centric dataset that may be collected, and can be easily generalized to different problem contexts, evaluation metrics, and modeling methods.

References


About the Authors

Balaji Padmanabhan received his B.Tech in Computer Science from the Indian Institute of Technology, Madras and Ph.D. in Information Systems from New York University. He is an assistant professor in the Operations and Information Management Department of The Wharton School, University of Pennsylvania. His research interests are in data mining with a focus on evaluation and applications in eCRM. His work has been published in leading conferences and journals both in Information Systems as well as in Computer Science.

Zhiqiang (Eric) Zheng is an assistant professor in Information Systems at the University of California, Riverside. He received his Ph.D. in Information Systems from the University of Pennsylvania. His research interests include data mining, information valuation and acquisition, and IT innovation and diffusion. His work has been published in MIS Quarterly, Management Science, INFORMS Journal on Computing, and Group Decision and Negotiation.

Steven O. Kimbrough received his MS (Industrial Engineering) and Ph.D. (Philosophy) degrees from the University of Wisconsin, Madison. He is a professor in the Operations and Information Management Department of The Wharton School, where he has been since 1984. His main research interests are in computational rationality (agents, games, and evolution); in logic modeling (especially agent communication languages and formal languages for business communication); and in text and text data mining.
Appendix A

Algorithm Probabilistic Clipping

Let $S_1, S_2, ..., S_n$ be $N$ user sessions in a site’s site-centric data. Assume that in this data the number of unique users is $M$ and users are identified by a userid $\{1,2, ..., M\}$. We define each session $S_i$ to be a tuple of the form $<u_i, C_i>$ where $u_i$ is the userid corresponding to the user in session $S_i$ and $C_i$ is a set of tuples of the form $<page, accessdetails>$, where each tuple represents data that a site captures on each user click. Corresponding to a click, $page$ is the page accessed and $accessdetails$ is a set of attribute-value pairs that represents any other information that a site can capture from each user click. This includes standard information from http headers such as time of access, IP address, referrer field, etc., and other information such as whether the user made a purchase in this page. For example, based on the above representation scheme, a user session at Expedia is represented as follows:

$$S_i = \{<time, IP, page, accessdetails> | \text{where } page, accessdetails, \text{and } IP \text{ are defined as above}\}$$

Given a session $S_i = <u_i, C_i>$, define function $kth-click(S_i, j)$ that returns a tuple $<page, accessdetails>$ if page is the $j^{th}$ page accessed in the session as determined from the time each page is accessed. For instance, in the above example $kth-click(S_i, 2)$ is $<flights.html, \{(time, 02/01/2001 23:45:15), (IP, 128.122.195.3)\}>$.

Also we define function $fragment(S_i, j, \text{fraglength})$ that represents data captured from a set of consecutive clicks in the session. In particular, $fragment(S_i, j, \text{fraglength})$ is the set of all tuples $kth-click(S_i, m)$ such that $j \leq m \leq \text{minimum}(j + \text{fraglength} -1, |C_i|)$, where $S_i = <u_i, C_i>$. For example, $fragment(S_i, 2, 2) = \{<flights.html, \{(time, 02/01/2001 23:45:15), (IP, 128.122.195.3)\}>, <hotels.html, \{(time, 02/01/2001 23:45:45), (IP, 128.122.195.3)\}>\}$. For any given set, $f$, of $<page, accessdetails>$ pairs, and any given session $S_i$, we say $f$ is a fragment of $S_i$ if there exists $j, k$ such that $f = fragment(S_i, j, k)$.

Finally, prior work (Mena 1999; Mobasher et al. 2000; VanderMeer et al. 2000) in building online customer interaction models assumes that three sets of variables are particularly relevant: (1) current visit summaries (e.g., time spent in current session), (2) historical summaries of the user (e.g., average time spent per session in the past), and (3) user demographics. Corresponding to this, we assume three user-defined functions as inputs to the algorithms:

1. $\text{summarize\_current}(f, S_i)$, defined when $f$ is a fragment of $S_i$. This function is assumed to return user-defined summary variables for the current fragment and session. For example, for the running example used in this section, $\text{summarize\_current}(\text{fragment}(S_i, 1, 2), S_i)$ may return $\text{numpages}=2, \text{tot\_time}=150$ seconds, $\text{booked}=1$ assuming the user made a booking in one of the three pages accessed in the session.

2. $\text{summarize\_historical}(f, S, i)$, where $f$ is a fragment of session $S_i$ and $S = \{S_1, S_2, ..., S_n\}$. This function is assumed to return summary variables based on all previous sessions. Note that the historical summaries are usually about the specific user in session $S_i$.

3. $\text{demographics}(u_i)$, which returns the demographic information available about user $u_i$.

Prior work in the IS and marketing literature proposed different sets of relevant usage metrics (Cutler 2000; Johnson et al. 2004; Kimbrough et al. 2000; Korgaonkar and Wolin 1999; Novak and Hoffman 1997). Borrowing on these for the experiments presented in this paper, we define $\text{summarize\_historical}, \text{summarize\_current}, \text{and demographics}(u_i)$ to return 15 metrics for site-centric data and 40 for user-centric data. These metrics are presented in Appendix B.

Algorithm probabilistic clipping is presented in Figure A1. Based on the desired data size, the sample rate is first computed in steps 1 through 7. Steps 9 through 25 iterate over all the sessions repeatedly until the desired number of records is sampled. Each time, a session is sampled probabilistically based on the expected number of records that should be derived from it.
Appendix B

Metrics for Within-Session Prediction

Table B1 details the metrics for the within-session prediction problem. For space considerations, we use the same table to explain the metrics for future-session and repeat-visit problems. Both problems include 30 independent variables: 12 site-centric ones (numbers 1 through 12 in the table) and 18 additional user-centric ones (numbers 16 through 33). Also, the dependent variables for these two problems are different: (1) for future-session if prediction, the dependent variable indicates a user will book in the future (after the session); (2) for repeat-visit prediction, it indicates if he will repeat visit in the future (after this session). Note that even though the two problems share the same metrics as within-session prediction, the values of metrics are generated from a different dataset as detailed in the “Methodology” section of the paper.
Table B1. Metrics for the Within-Session Prediction Problem

<table>
<thead>
<tr>
<th>No.</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gender</td>
<td>“1”—Male, “0”—Female</td>
</tr>
<tr>
<td>2</td>
<td>age</td>
<td>Age of the user</td>
</tr>
<tr>
<td>3</td>
<td>income</td>
<td>Income of the user</td>
</tr>
<tr>
<td>4</td>
<td>edu</td>
<td>“0”—high school or less, “1”—college, “2”—post college</td>
</tr>
<tr>
<td>5</td>
<td>hhsize</td>
<td>Size of house hold</td>
</tr>
<tr>
<td>6</td>
<td>child</td>
<td>“1”—have, “0”—not have</td>
</tr>
<tr>
<td>7</td>
<td>booklh</td>
<td>No. of bookings the user made at this site in the past</td>
</tr>
<tr>
<td>8</td>
<td>sesslh</td>
<td>No. of sessions to this site so far</td>
</tr>
<tr>
<td>9</td>
<td>freqlh</td>
<td>A measure of the frequency of purchases to this site, defined as booklh/sesslh</td>
</tr>
<tr>
<td>10</td>
<td>minutelh</td>
<td>Time spent in this site so far in minutes</td>
</tr>
<tr>
<td>11</td>
<td>hpsesslh</td>
<td>Average hits per session to this site</td>
</tr>
<tr>
<td>12</td>
<td>mpsesslh</td>
<td>Average time spent per sessions to this site</td>
</tr>
<tr>
<td>13</td>
<td>hitlc</td>
<td>No. of hits to this site up to this point in this session</td>
</tr>
<tr>
<td>14</td>
<td>minutelc</td>
<td>Time spent up to this point in this session</td>
</tr>
<tr>
<td>15</td>
<td>weekend</td>
<td>Indicating if this session occurs on weekend</td>
</tr>
<tr>
<td>16</td>
<td>bookgh</td>
<td>No. of past bookings of all sites so far</td>
</tr>
<tr>
<td>17</td>
<td>sespsite</td>
<td>Average sessions per site so far</td>
</tr>
<tr>
<td>18</td>
<td>sessgh</td>
<td>Total no. of sessions visited of all sites so far</td>
</tr>
<tr>
<td>19</td>
<td>freqgh</td>
<td>A measure of the frequency of purchases across all sites, defined as bookgh/sessgh</td>
</tr>
<tr>
<td>20</td>
<td>minutegh</td>
<td>Total minutes of all sites</td>
</tr>
<tr>
<td>21</td>
<td>hpsessgh</td>
<td>Average hits per session</td>
</tr>
<tr>
<td>22</td>
<td>mpsessgh</td>
<td>Average minute per session</td>
</tr>
<tr>
<td>23</td>
<td>awareset</td>
<td>Total no. of unique shopping sites visited</td>
</tr>
<tr>
<td>24</td>
<td>basket</td>
<td>Average no. of shopping sites visited per session</td>
</tr>
<tr>
<td>25</td>
<td>single</td>
<td>Percentage of single-site sessions</td>
</tr>
<tr>
<td>26</td>
<td>booksh</td>
<td>Percentage of total bookings are to this site</td>
</tr>
<tr>
<td>27</td>
<td>hitsh</td>
<td>Percentage of total hits are to this site</td>
</tr>
<tr>
<td>28</td>
<td>sesssh</td>
<td>Percentage of total sessions are to this site</td>
</tr>
<tr>
<td>29</td>
<td>minutesh</td>
<td>Percentage of total minutes are to this site</td>
</tr>
<tr>
<td>30</td>
<td>entrate</td>
<td>No. of sessions start with this site/total sessions of this site</td>
</tr>
<tr>
<td>31</td>
<td>peakrate</td>
<td>No. of sessions the user spend the most time within this site/total sessions</td>
</tr>
<tr>
<td>32</td>
<td>exitrate</td>
<td>No. of sessions end with this site/total sessions of this site</td>
</tr>
<tr>
<td>33</td>
<td>SERate</td>
<td>No. of sessions start with this site/total sessions of this site</td>
</tr>
<tr>
<td>34</td>
<td>hitgc</td>
<td>Total hits of all sites in the current session</td>
</tr>
<tr>
<td>35</td>
<td>basketgc</td>
<td>No. of shopping sites in this session</td>
</tr>
<tr>
<td>36</td>
<td>minutegc</td>
<td>Time spent of all sites in this session</td>
</tr>
<tr>
<td>37</td>
<td>SEgc</td>
<td>Indicating if this session uses search engines</td>
</tr>
<tr>
<td>38</td>
<td>path</td>
<td>Indicating if this site is an entry/peak</td>
</tr>
<tr>
<td>39</td>
<td>hitshc</td>
<td>Hits to this site/ hits to all sites in this session</td>
</tr>
<tr>
<td>40</td>
<td>minutshc</td>
<td>Minutes to this site/total minutes in this session</td>
</tr>
<tr>
<td>41</td>
<td>bookfut</td>
<td>Binary dependent variable, indicating if this user is going to book in the remainder of the session (after the clipping point)</td>
</tr>
</tbody>
</table>

Note: Variables 1 through 15 are site-centric variables; 16 through 40 are additional user-centric variables; variable 41 is the dependent variable.
Appendix C

Rationale Behind Variable Construction

The user-centric data provided to us consisted of household ID (HHID), the time a user visits each page, the duration of the visit, the site visited and the URL of the specific page. Each site is also pre-categorized into categories such as search engine, travel, etc. In order to develop a set of variables from the raw usage data, we used a framework within which these variables can be identified. A central concept in the framework is that of a session, which is a collection of hits (or clicks) by a single user during a single browsing period. For each session, the data is of the format (page, time, domain), (page2, time2, domain2). Given this structure, for the current session we created features with respect to hit (page), time, and site (domain). As Table C1 shows, features related to hits and time are identically created for both site-centric and user-centric datasets. With respect to domain, all site-centric sessions are under the same domain, while user-centric sessions consist of pages from different domains. Hence for user-centric datasets, we construct three types of cross-domain features that do not have an equivalent in site-centric data:

1. Path-related features that capture the time sequence of visits to different sites. Three important characteristics of a path are where the session starts, peaks, and ends.

2. User-centric data also tells us information about what sites a user is visiting. We note that two types of sites may be particularly informative about shopping behavior. The use of search engines is important since it suggests a focused need for the user. Visiting shopping sites is again relevant since it broadly indicates interest in things that can be bought. Hence under site we identified two such variables for user-centric data.

3. User-centric data also has related market share information. Specifically, we construct time and hit related market share metrics.

Similarly, for the historical sessions, we also have information on hit, time, and site. From the site-centric data, we create four variables: sesslh, minutelh, hpsesslh, and mpsesslh (see Table C1). From user-centric data, in addition, we know the share of a user’s activities to this site, the path, and the site characteristics of the past sessions. Thus we created three metrics (sessh, minuteshc, hitshc) describing the share of activities; another three metrics (entratre, peakrate, and exitrate) describing the path, and five metrics (SErate, basket, single, awareset, and sespsite) describing the characteristics of sites. Finally, past purchases are important indicators of future purchases. From site-centric data, we have booklh (total number of purchases to this site) and freqlh (percentage of user sessions in which a booking was made at this site). From user-centric data, we know bookgh (total number of purchases across sites), freqgh (percentage of user sessions in which a booking was made at any site), and booksh (the share of purchases to this site). Table C1 details the site-centric and user-centric metrics developed under the framework.

| Table C1. Identified Site-Centric and User-Centric Metrics Under the Framework |
|-----------------------------|-----------------------------|-----------------------------|
| Granularity Level         | Statistics On              | User-Centric | Site-Centric |
| Current Session           | Hit                         | Hitgc          | hitlc         |
|                           | Time                        | Minutelc, weekend |
|                           | Path                        | Mutselc, weekend |
|                           | Sites                       | basketgc, SEgc   |
|                           | Share                       | minuteshc, hitshc  |
|                           | Hit                         | Hpsessgh       |
|                           | Sessions                    | sesslh          |
|                           | Time                        | Minutelgh, mpsessgh|
| All past sessions         | Path                        | entratre, peakrate, exitrate |
|                           | Share                       | sessh, miutesh, hitsh |
|                           | Sites                       | SErate, single, basket, awareset, sespsite |
|                           | Shopping                    | bookgh, freqgh, booksh  |
| User                      | Demographics                | 6 variables                |

A common rule of thumb is to observe the time associated with each hit and group consecutive hits that are within 30 minutes of each other into a session (http://w3.org/WCA/1999/01/Terms.html).